IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rahin Mansoor Ali Al-Hammadi
08/22/2023

Outline

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory data analysis (EDA) with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- EDA results
- Interactive analytics
- Predictive analysis

Introduction

Project background and context

SpaceX (Space Exploration Technologies Corporation) is a space transportation and aerospace manufacturer founded in 2002 by Elon Musk.

SpaceX has been a disruptive force in the worldwide launch industry as its launch services are less expensive than many of its competitors.

SpaceX charges $62 million for a Falcon 9 rocket launch, much cheaper than competitor startups like the European launcher Arianespace's Ariane 5 or U.S. rocket builder United Launch Alliance’s (ULA) Atlas V, which can cost up to $165 million.

Problems you want to find answers

- The aim of this project is predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully based on the data available.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

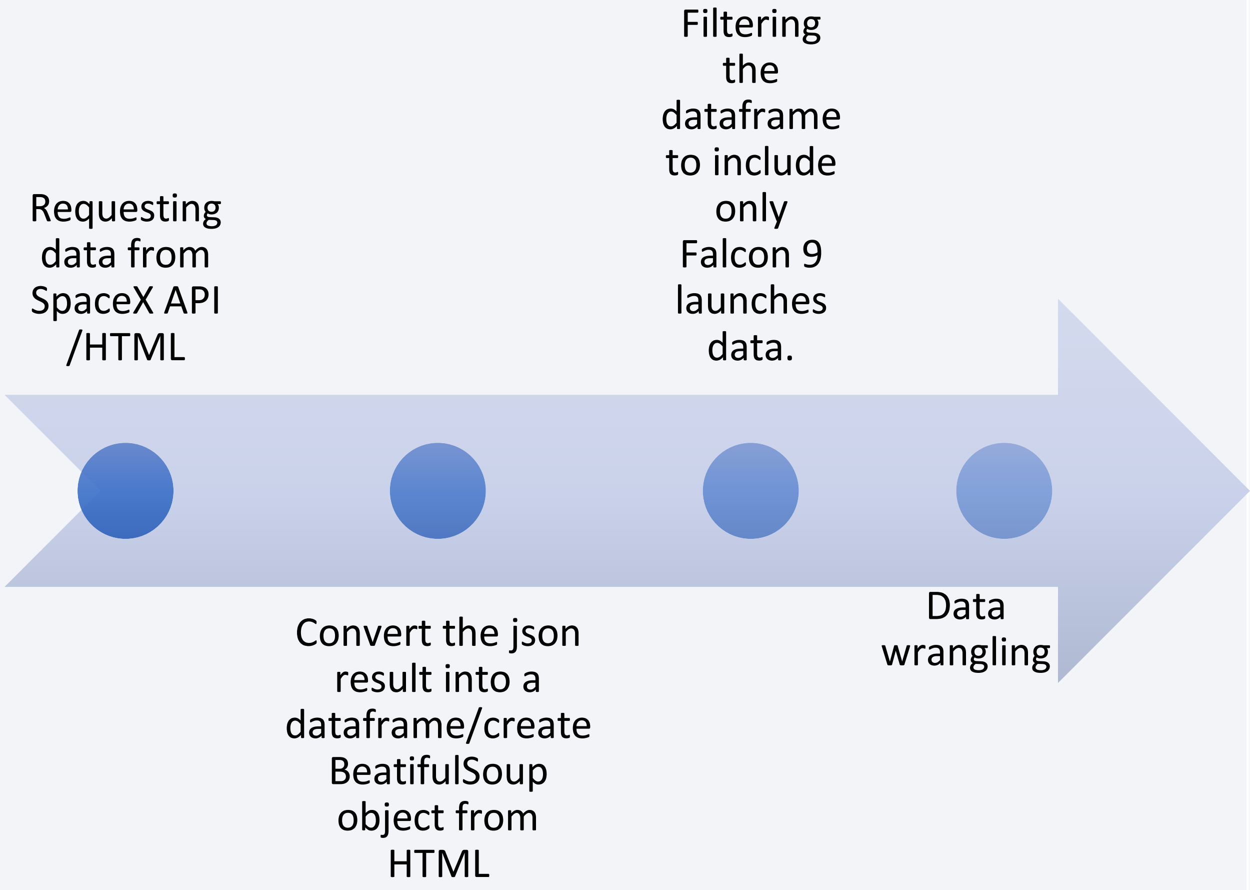
The data sets were collected:

•Requesting rocket launch data from SpaceX API or from HTML.

•Return output will be the SpaceX data in JSON format. We apply the json_normalize function to convert the json result into a dataframe. In Case using webscraping to collect the data we applies BeautifulSoup object from the HTML response.

•Filtering the dataframe to include only Falcon 9 launches data.

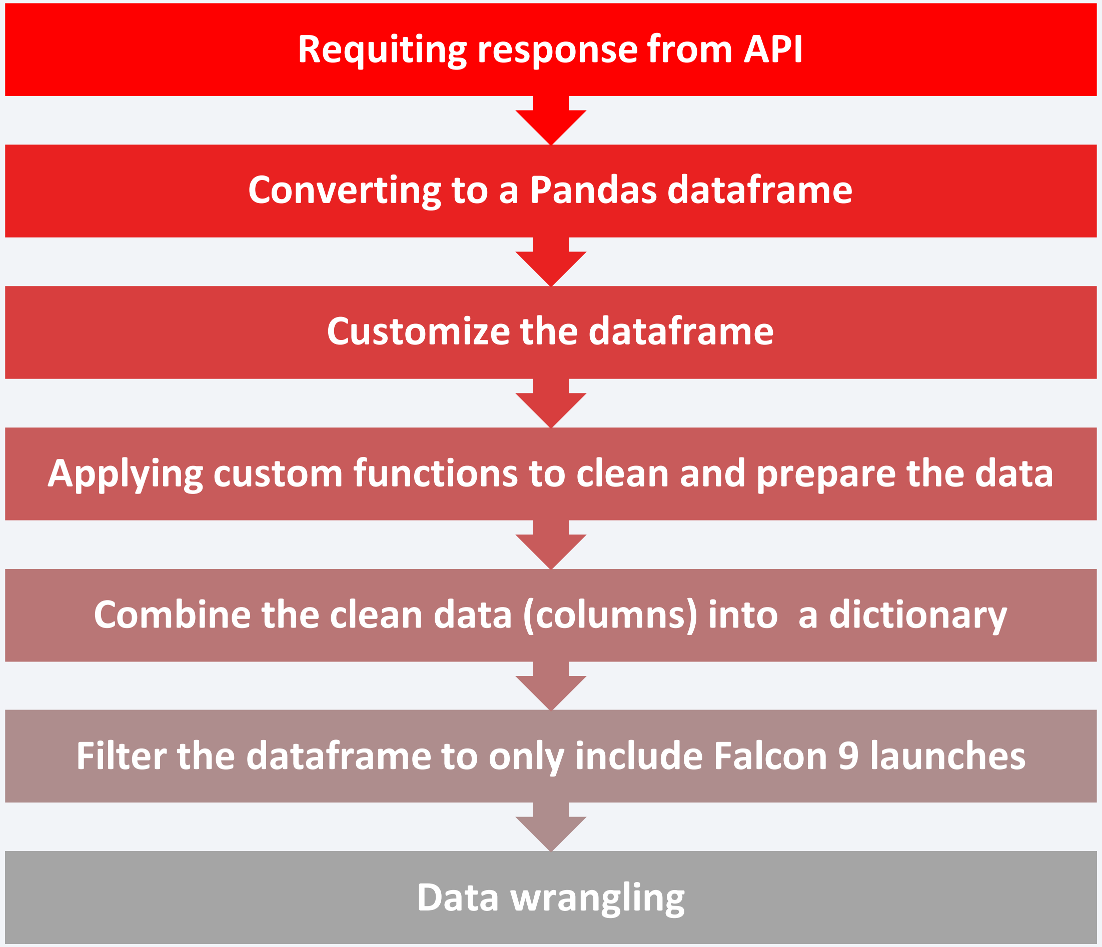
•Data wrangling.



Data Collection – SpaceX API

Data collection with SpaceX REST:

1. Requiting response from API
2. Converting the data from JSON format to a Pandas dataframe using .json_normalize()
3. Customize the dataframe (removing columns that we are not using, convert the date format, etc.)
4. Applying custom functions to clean and prepare the data
5. Combine the clean data (columns) into a dictionary
6. Filter the dataframe to only include Falcon 9 launches
7. Data wrangling

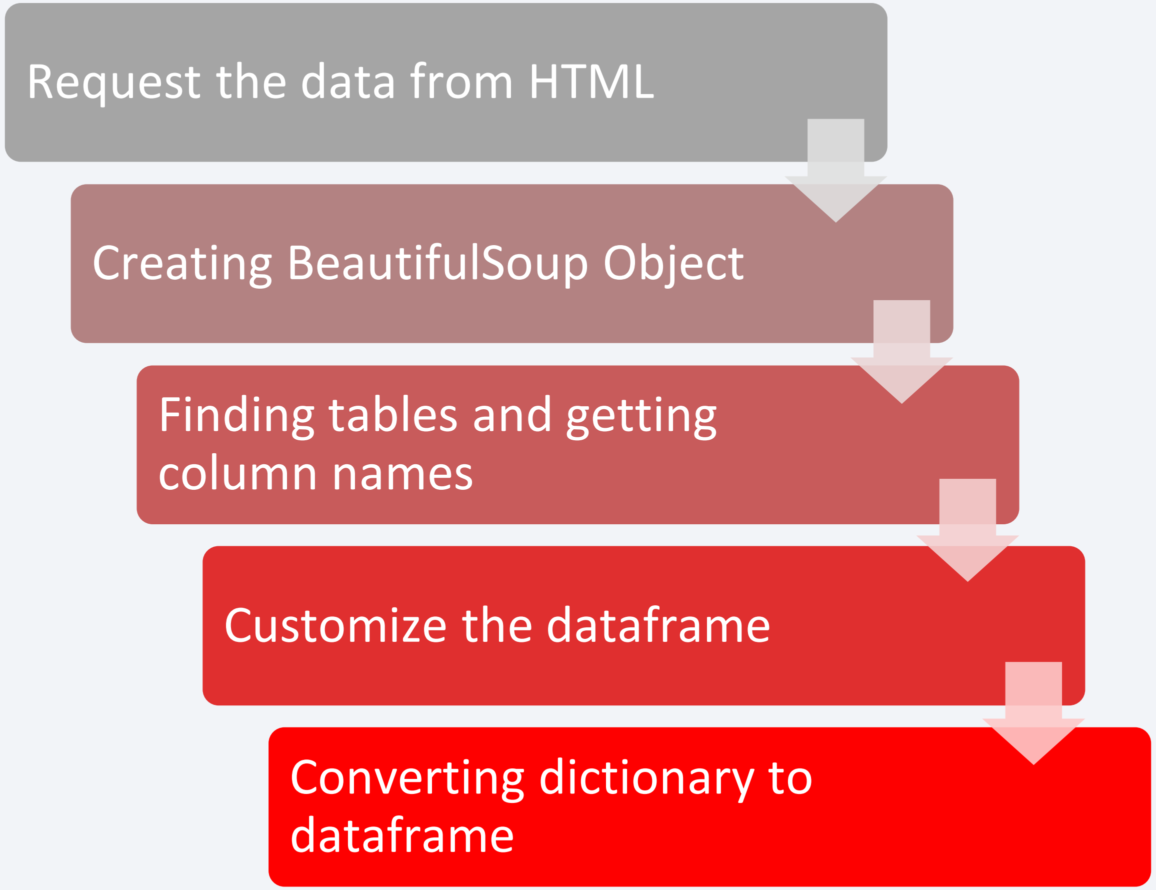


https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Data collection with web scraping

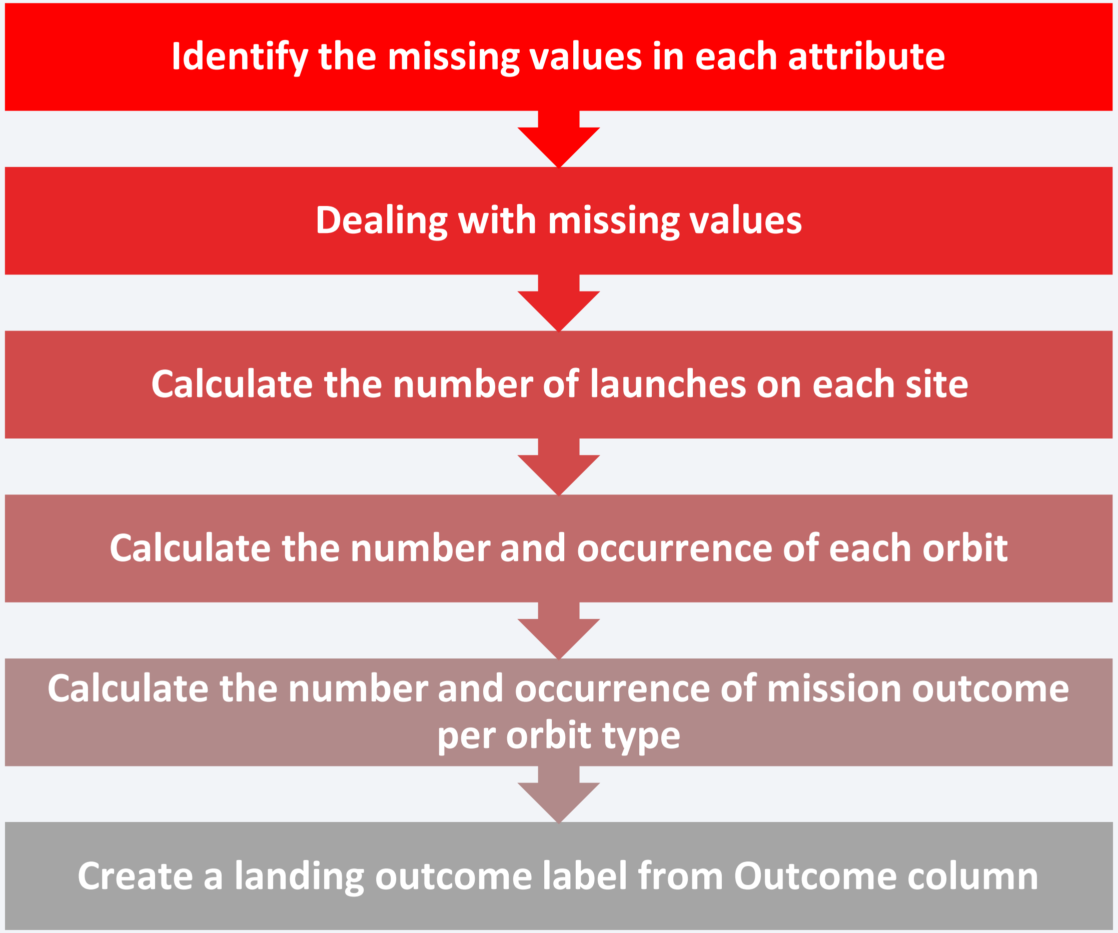
1. Request the Falcon9 Launch Wiki page from its URL as HTML
2. Create a BeautifulSoup object from the HTML response
3. Extract all column/variable names from the HTML table header
4. Customize the dataframe (create an empty dictionary with keys from the extracted column names, fill up the launch_dict with launch records extracted from table rows, clean and prepare the data)
5. Create a data frame by parsing the launch HTML tables



https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Identify the missing values in each attribute
- Dealing with missing values in PayloadMass attribute.
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column



https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-spacex-data_wrangling.ipynb

EDA with SQL

- Download and Connect to the database
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Download and Connect to the database

Display the some data (CCA, CRS) do some colleclation (average)

List and filtering the date (e.g., total number of successful and failure mission outcomes)

Rank the count of successful landing_outcomes

https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-SQL.ipynb

EDA with Data Visualization

Scatter plot	• the FlightNumber vs. PayloadMassand overlay the outcome of the launch.
Scatter plot	• the relationship between Flight Number and Launch Site
Bar plot	• the relationship between Payload and Launch Site
Scatter plot	• the relationship between success rate of each orbit type
Scatter plot	• the relationship between FlightNumber and Orbit type
Scatter plot	• the relationship between Payload and Orbit type
Line plot	• the launch success yearly trend

https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-data-VIZ.ipynb

Build an Interactive Map with Folium

Mark all launch msites on a map

Mark the success/failed launches for each site on the map

Calculate the distances between a launch site to its proximities

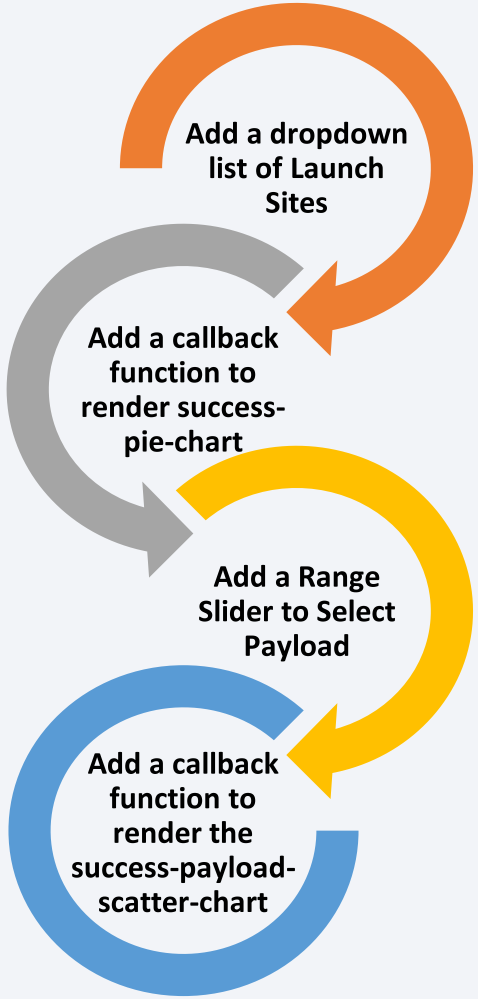
By adding those markers:

- It enable us to find some geographical patterns about launch sites.
- It enable us to easily identify which launch sites have relatively high success rates.

https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-lab_launch_site_location-map-Folium.ipynb

Build a Dashboard with Plotly Dash

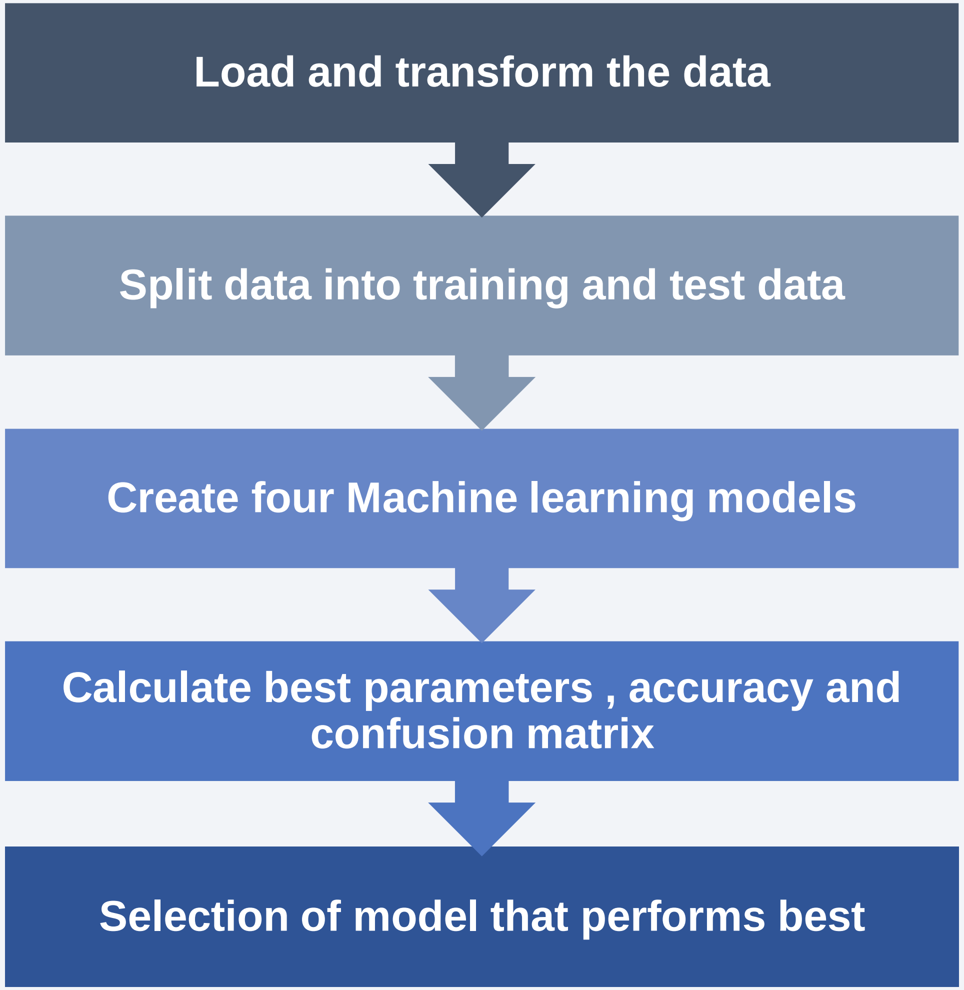
- Add a Launch Site Drop-down Input Component
- Add a callback function to render success-pie-chart based on selected site dropdown
- Add a Range Slider to Select Payload
- Add a callback function to render the success-payload-scatter-chart scatter plot



https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-Dashboard.ipynb

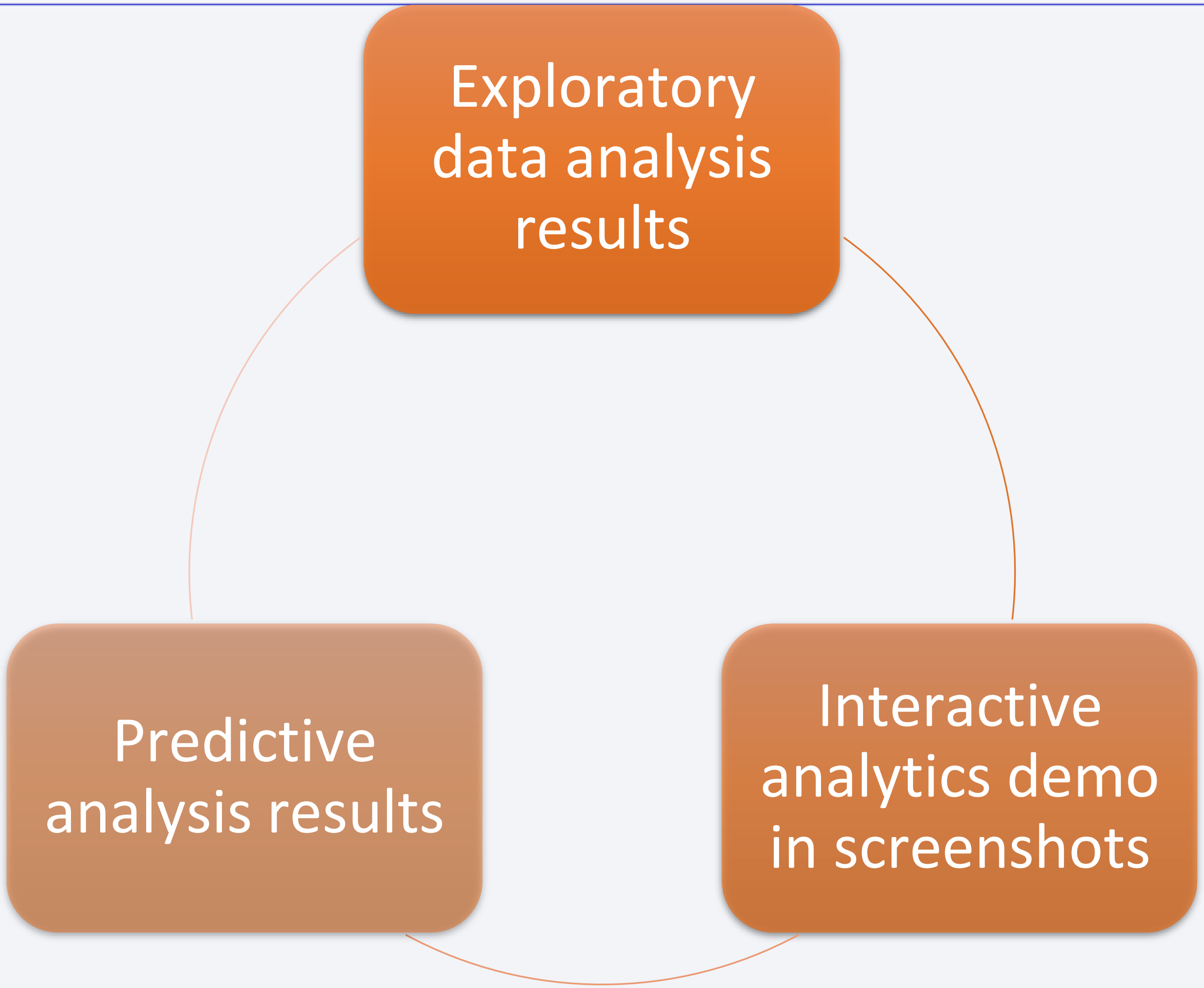
Predictive Analysis (Classification)

- Load and transform the data
- Train_test_split to split data into training and test data.
- Create four Machine learning models (Logistic regression, SVM, Decision Tree, KNN), and create GridSearchCV objectives for each models.
- Calculate best parameters , accuracy and confusion matrix
- Selection of model that performs best which is DecisionTree with a score of 0.90.



https://github.com/Rahinmans/space-x-project-/blob/main/jupyter-labs-SpaceX_Machine_Learning_Prediction.ipynb

Results

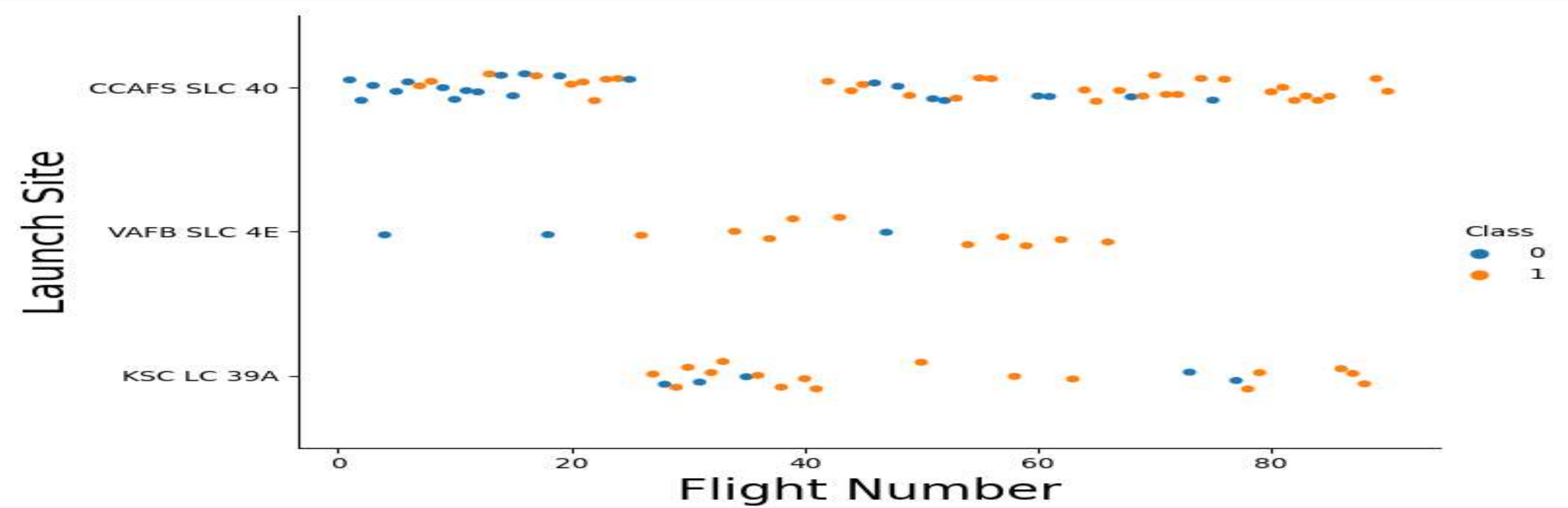

Exploratory data analysis results
Interactive analytics demo in screenshots
Predictive analysis results

Section 2

Insights drawn from EDA

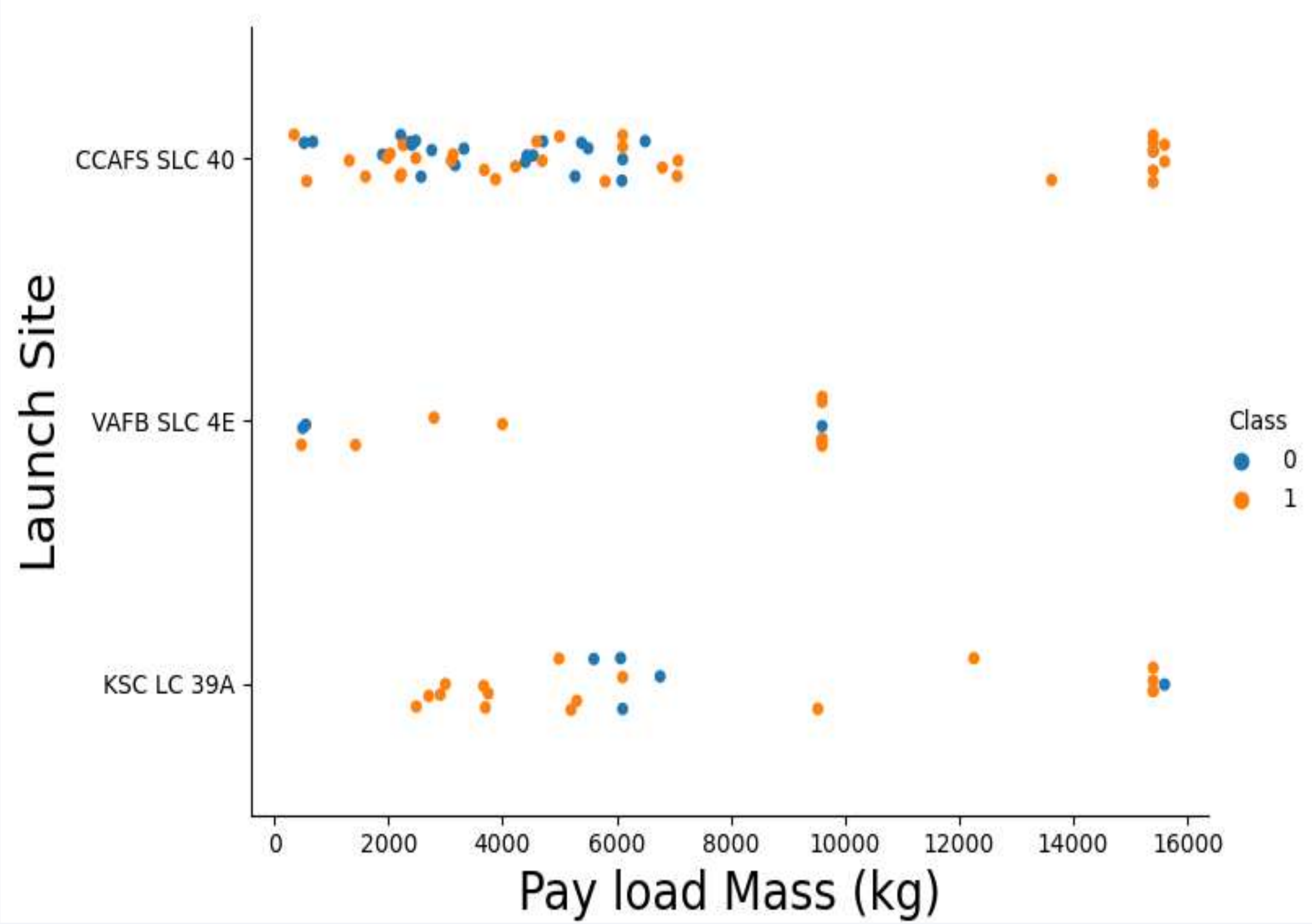
Flight Number vs. Launch Site

- With time the successful rate has increased for every Launch Site, particularly for the site “**CCAFS SLC 40**”, where are concentrated the majority of the launches.
- For VAFB SLC 4E and KSC LC 39A has a higher successful rate but represents one third of the total flight launches.



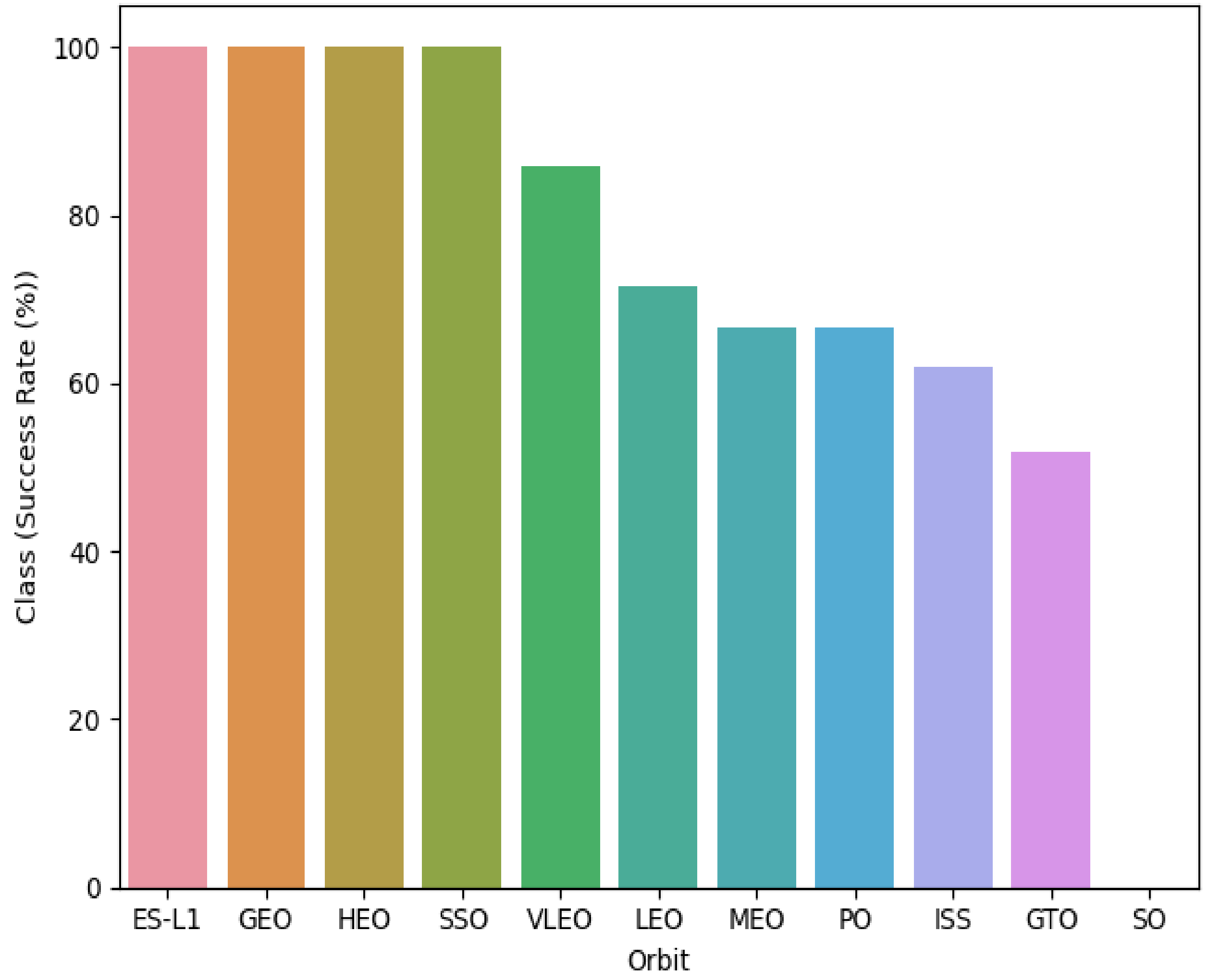
Payload vs. Launch Site

- In **VAFB SLC 4E** launch site there are no rockets launched for heavy payload mass (greater than 10000kg)
- In **KSC LC 39A** launch site there are no rockets launched for lower payload mass (less than 2500kg)
- In **CCAFS SLC 40** launch site the rockets launched with payload mass less than 7500kg and more than 13000kg, but not in between.



Success Rate vs. Orbit Type

- The first 4 orbit types (ES-L1, GEO, HEO, SSO) had the best successful rate.
- The bar chart must be interpreted with number of launches per Orbit type.

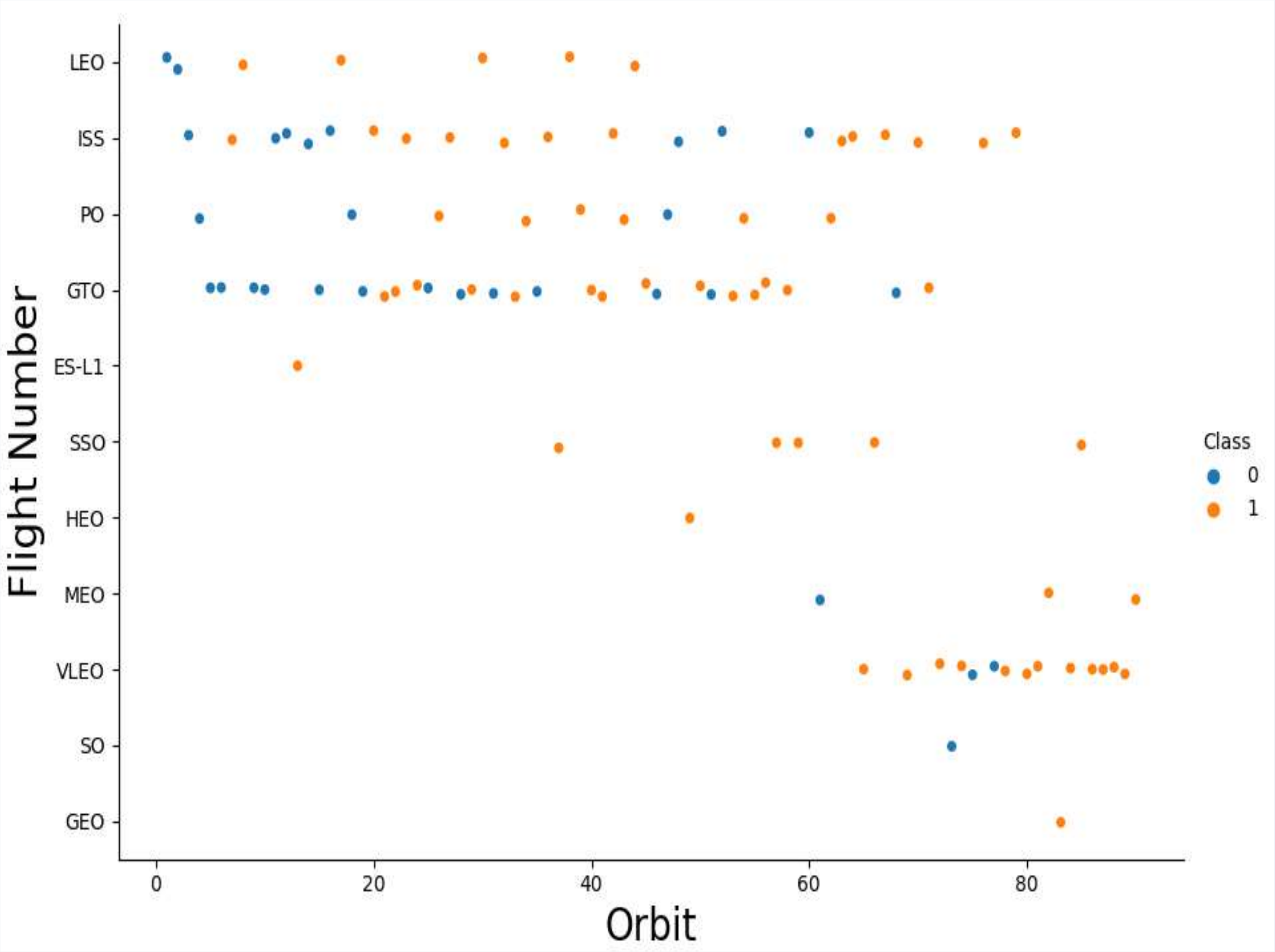


Flight Number vs. Orbit Type

- As expected, there are more failures at the beginning of the series of launches, but after the first 40 launches the ratio improves by reducing the 50% of unsuccessful landings.

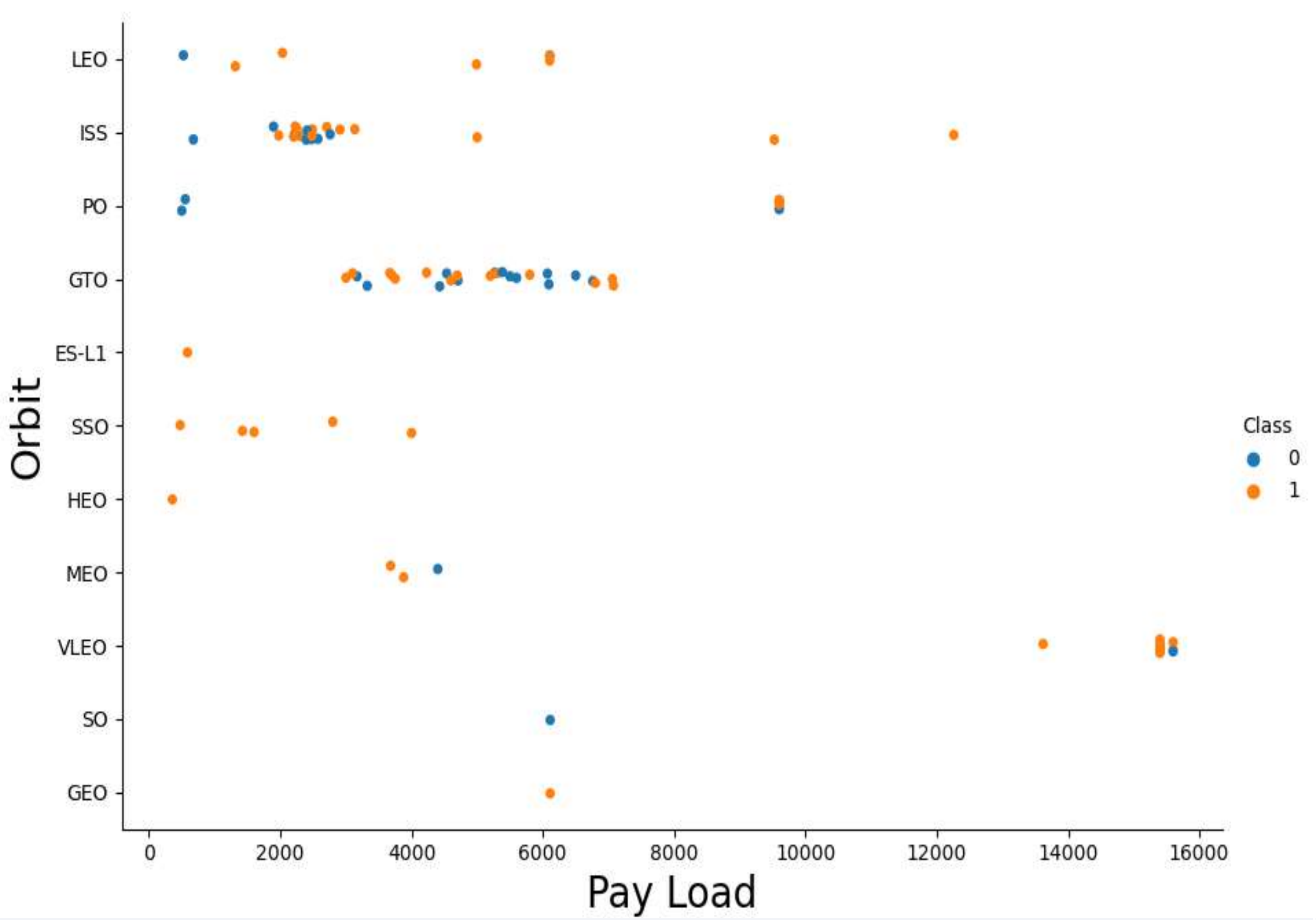
- GTO and ISS orbits has the higher concentration of launches with the lowest ration of successful landings.

- The orbits with higher successful rate, has one or just a few number of launches.



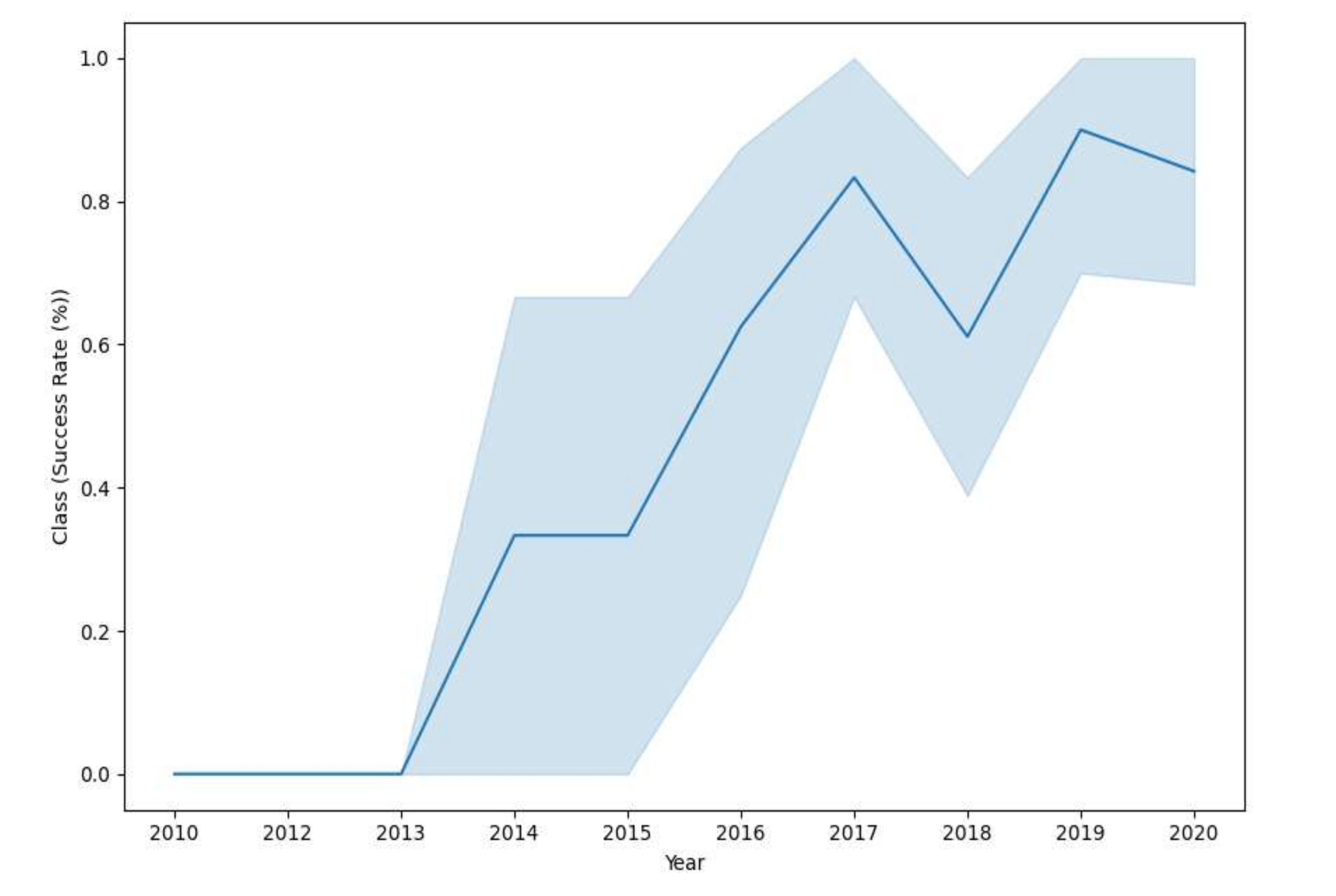
Payload vs. Orbit Type

- Exists a visible limit of Payload around 7600kg less than 10 launches exceed that limit.
- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- GTO, it can not find a different between successful or unsuccessful landings as the distribution here is lined and indicated no different in the payload volume.



Launch Success Yearly Trend

- The success rate of landings could be grouped into two groups:
  - For the first period (2010-2012) the success rate were very low.
  - For the second period (2013-2020) the success rate were gradually increased, although there was a slide drop in 2018.



All Launch Site Names

- The ‘DISTINCT’ function has been used to find the unique values in the launch site column.
- The four unique launch sites in the space mission:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

In [8]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[8]:

Launch_Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Using the query WHERE, LIKE, and LIMIT to get 5 records where launch sites begin with `CCA`

```
In [9]: %sql SELECT * FROM 'SPACEXTBL' WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

 * sqlite:///my_data1.db
Done.

Out[9]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Using SUM function and WHERE clause to calculate the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

In [10]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[10]:

Total Payload Mass(Kgs)	Customer
45596.0	NASA (CRS)

Average Payload Mass by F9 v1.1

- Using AVG() function and WHERE clause to calculate the average payload mass carried by booster version F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
In [11]:  %sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LI
```

```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

- Using MIN function and WHERE clause to find the dates of the first successful landing outcome on ground pad

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [20]:
```
%sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing_Outcome" = "Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[20]:

MIN(DATE)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using the WHERE clause and AND operator to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [22]:
```
%sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND PAYLOAD_MAS
```

```
 * sqlite:///my_data1.db
Done.
```

Out[22]:

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Using COUNT function with GROUP BY statement to calculate the total number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

In [23]:
```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

Mission_Outcome	Total
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Using subquery to find list the names of the booster which have carried the maximum payload mass

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
In [16]: %sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MA
```

```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600.0
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600.0
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600.0
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600.0
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600.0
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600.0
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600.0
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600.0
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600.0
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600.0
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600.0
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600.0

2015 Launch Records

- Using WHERE clause, LIKE to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
In [25]:  %sql SELECT substr(Date,7,4), substr(Date, 4, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Out
```

```
 * sqlite:///my_data1.db
Done.
```

Out[25]:

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing_Outcome
2015	10	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395.0	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898.0	Success	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Using COUNT function, WHERE clause BETWEEN operator and GROUP BY statement to rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[74]: %sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTBL WHERE "Landing_Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010'
```

```
 * sqlite:///my_data1.db
Done.
```

[74]:

Landing_Outcome	COUNT(Landing_Outcome)
Success (ground pad)	7
Success	20
Success (drone ship)	8

Section 3

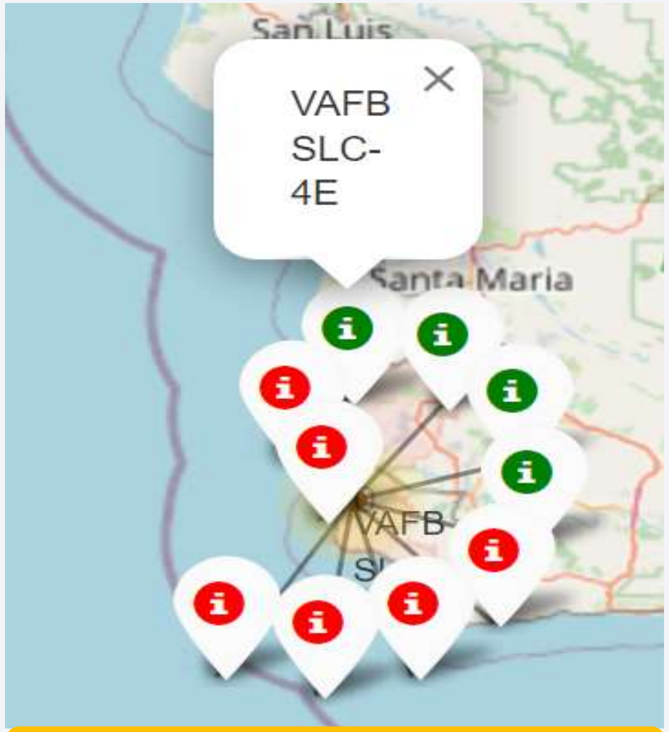
Launch Sites Proximities Analysis

All launch sites markers based on Global map

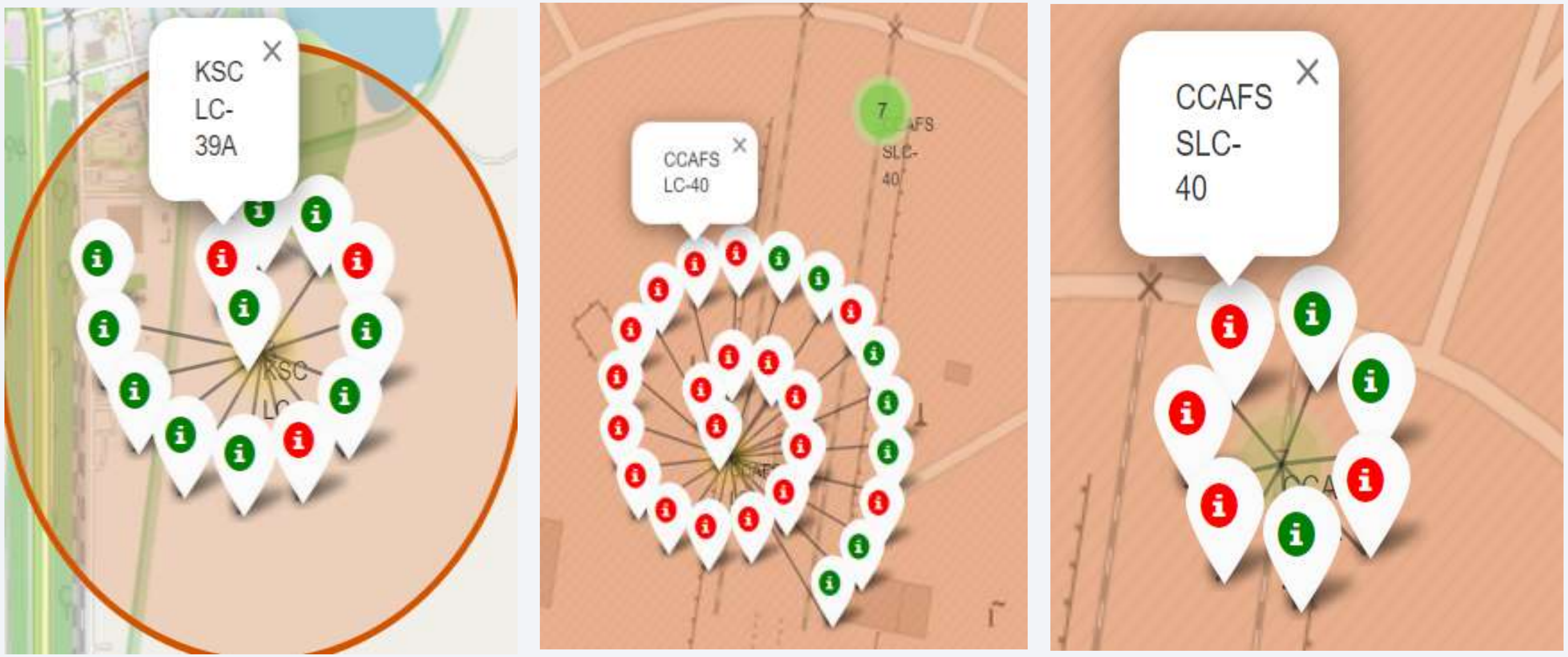
- The SpaceX launch sites are in the USA (Florida and California).

Launch sites with markers of successful launches

- The 3 launch sites Florida (KSC LC- 39A, CCAFS LC-40, CCAFS SLC-40) carried majority of launches.
- Based on the green markers which indicates the successful launch the **KSC LC- 39A** has the highest green markers compare to the red markers 10:3.



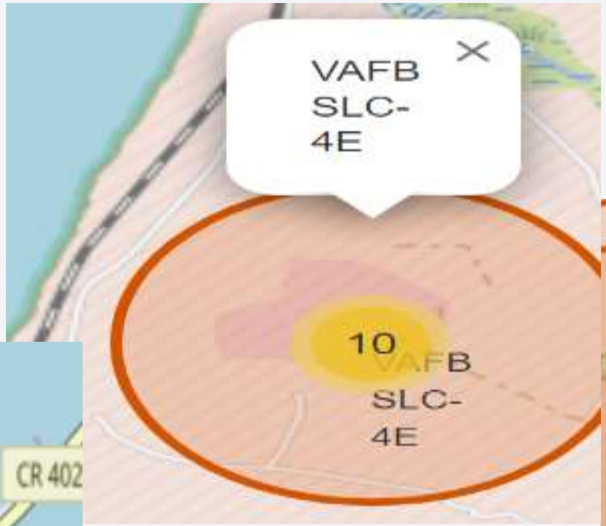
California



Florida

Launch sites with distance to landmarkes

- Approximately all the 4 sites to coastline and the closest site is CCAFS SLC-40 with 0.86KM to coastline.
- Approximately all sites keep away from cities.
- Approximately all sites close to railway and highway.

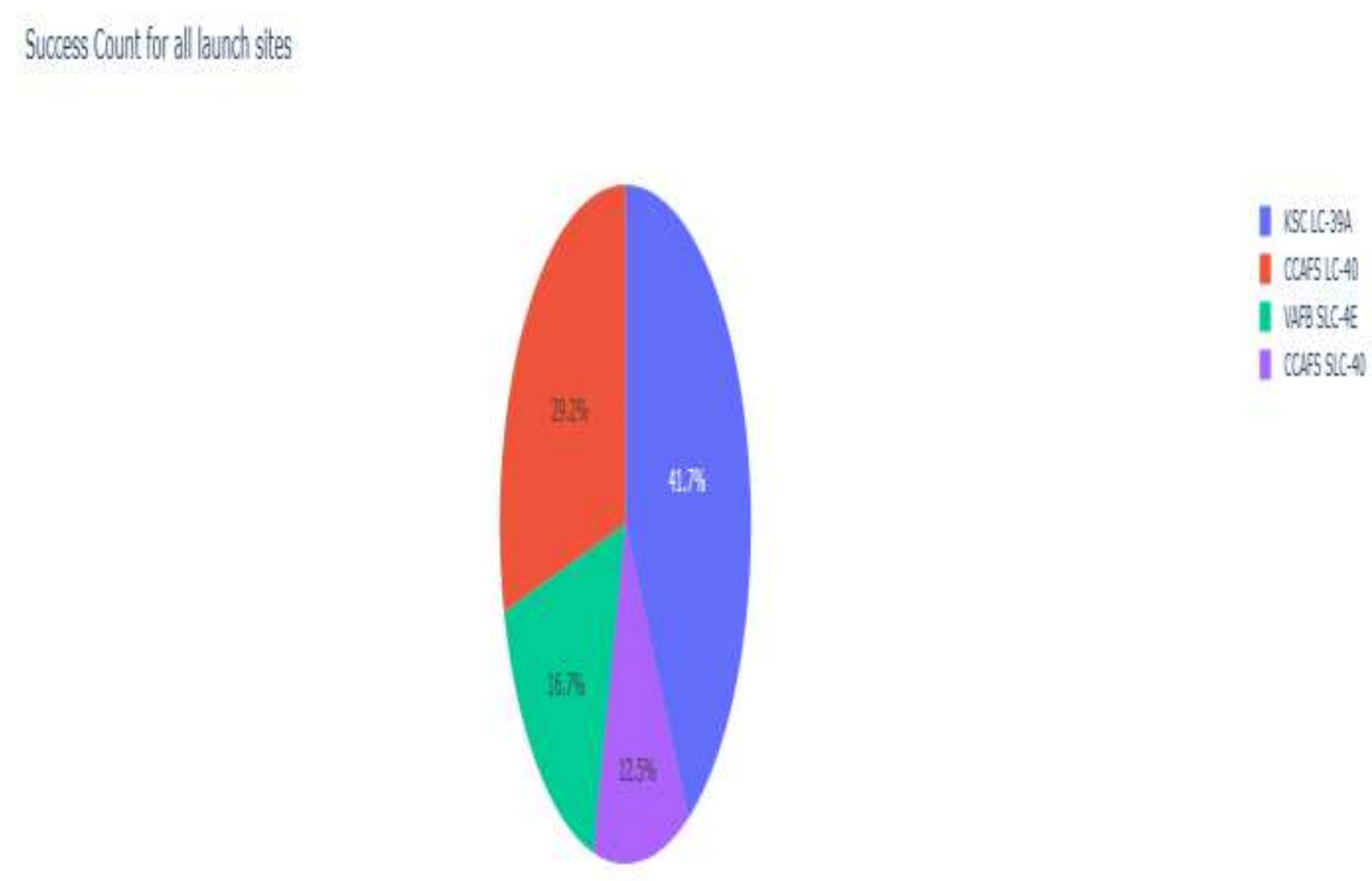


Section 4

Build a Dashboard with Plotly Dash

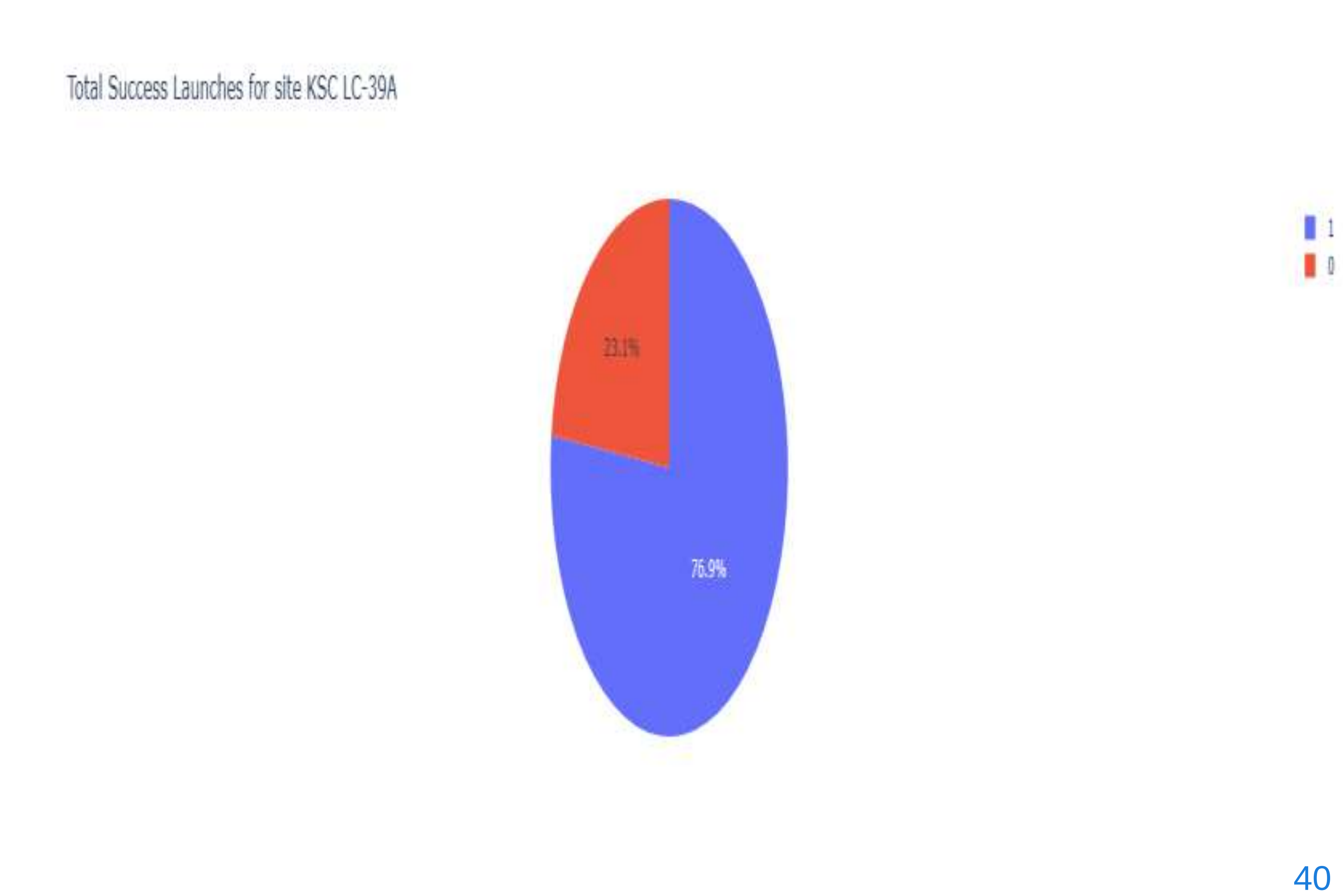
The success percentage for all launch sites

- The pie chart shows that 42% of the success launch was on the site (**KSC LC- 39A**) followed by (**CCAFS LC- 40**) with 29%.



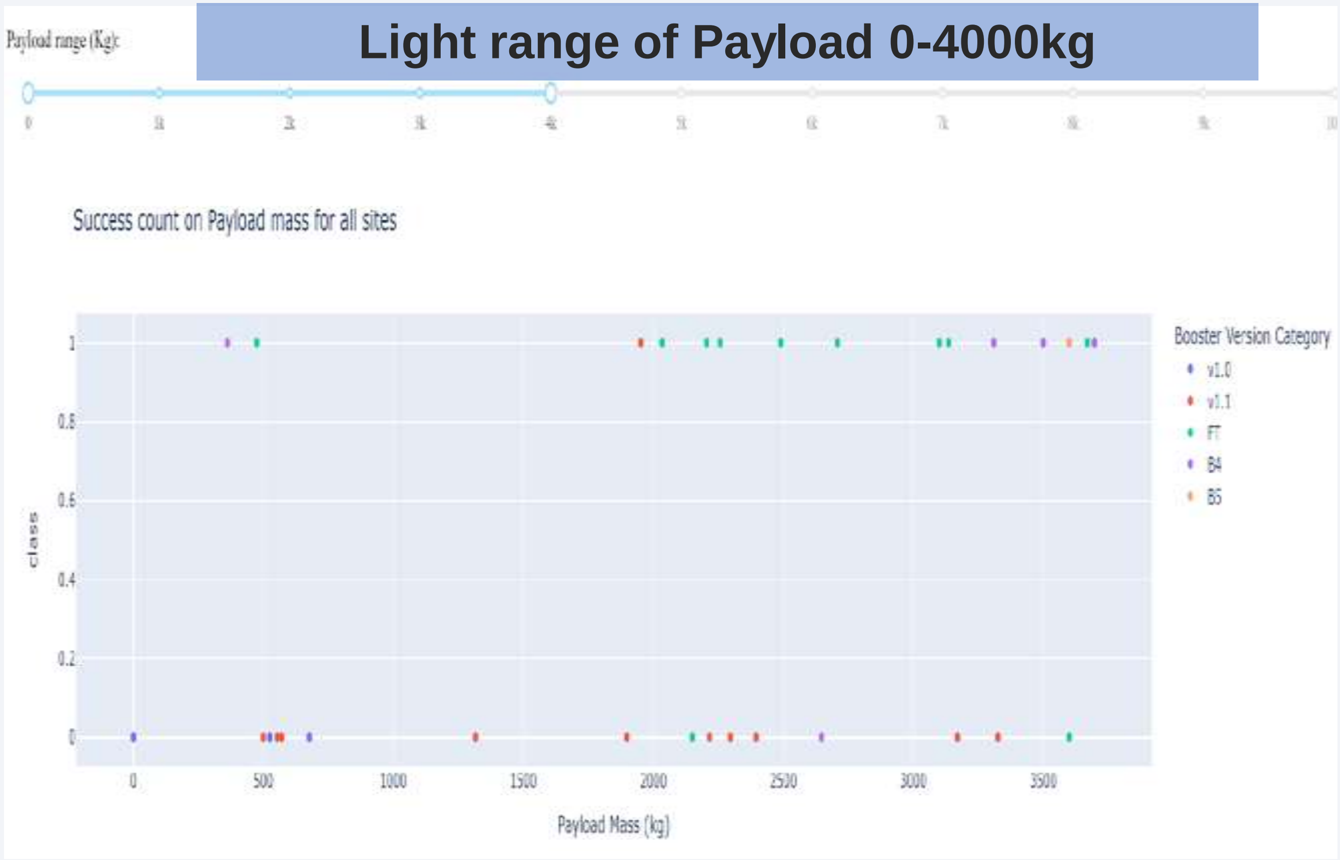
The launch site with the **highest launch success rate**

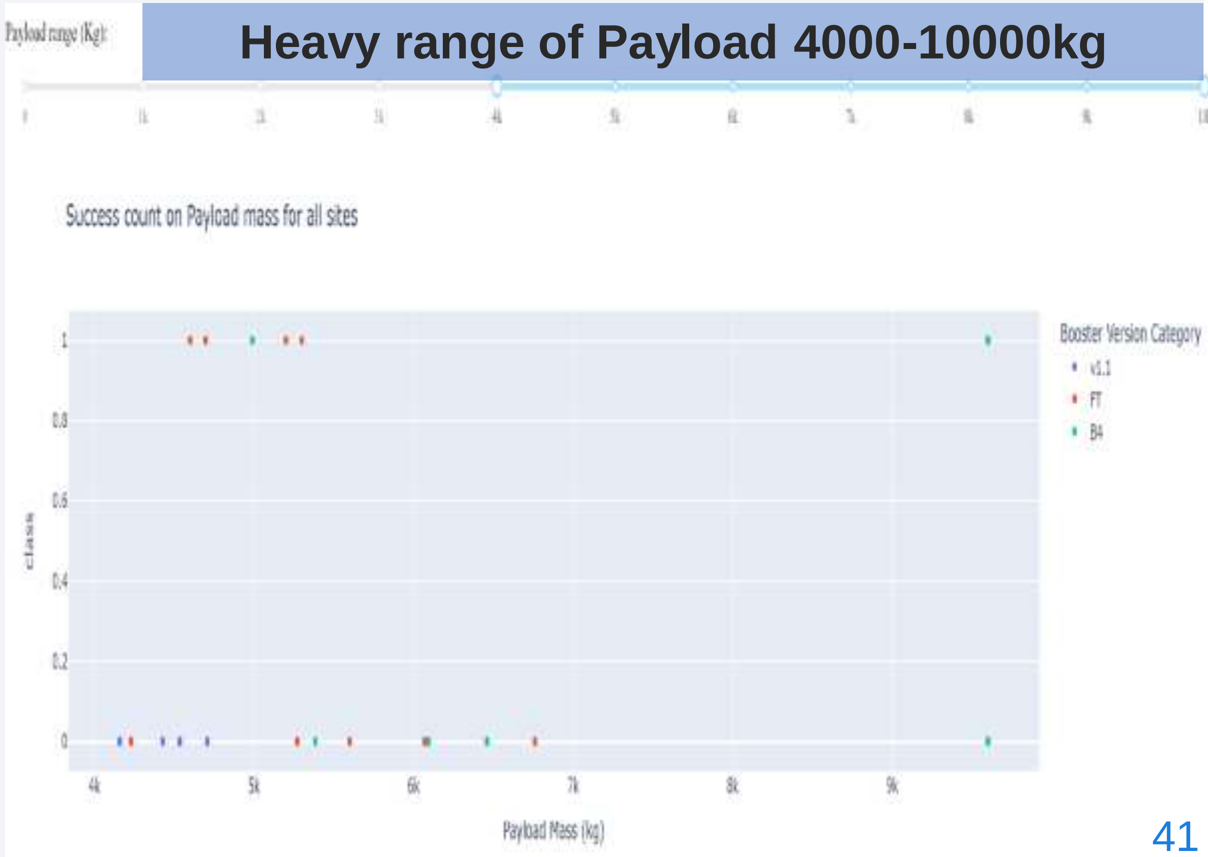
- In line with previous slide here is the highest total success launch rate which is in the site (**KSC LC-39A**) with 77% success launches.



Payload vs Launch outcome for all sites

- In the low range of payload range (0-4000kg) with booster version category (v1.0, v1.1, FT, B4, B5).
- In the heavy of payload range (4000-10000kg) with booster version category (v1.1, FT, B4).
- It is noticed that the FT booster version category has the highest success rate specially in the payload range of 2000-6000kg.



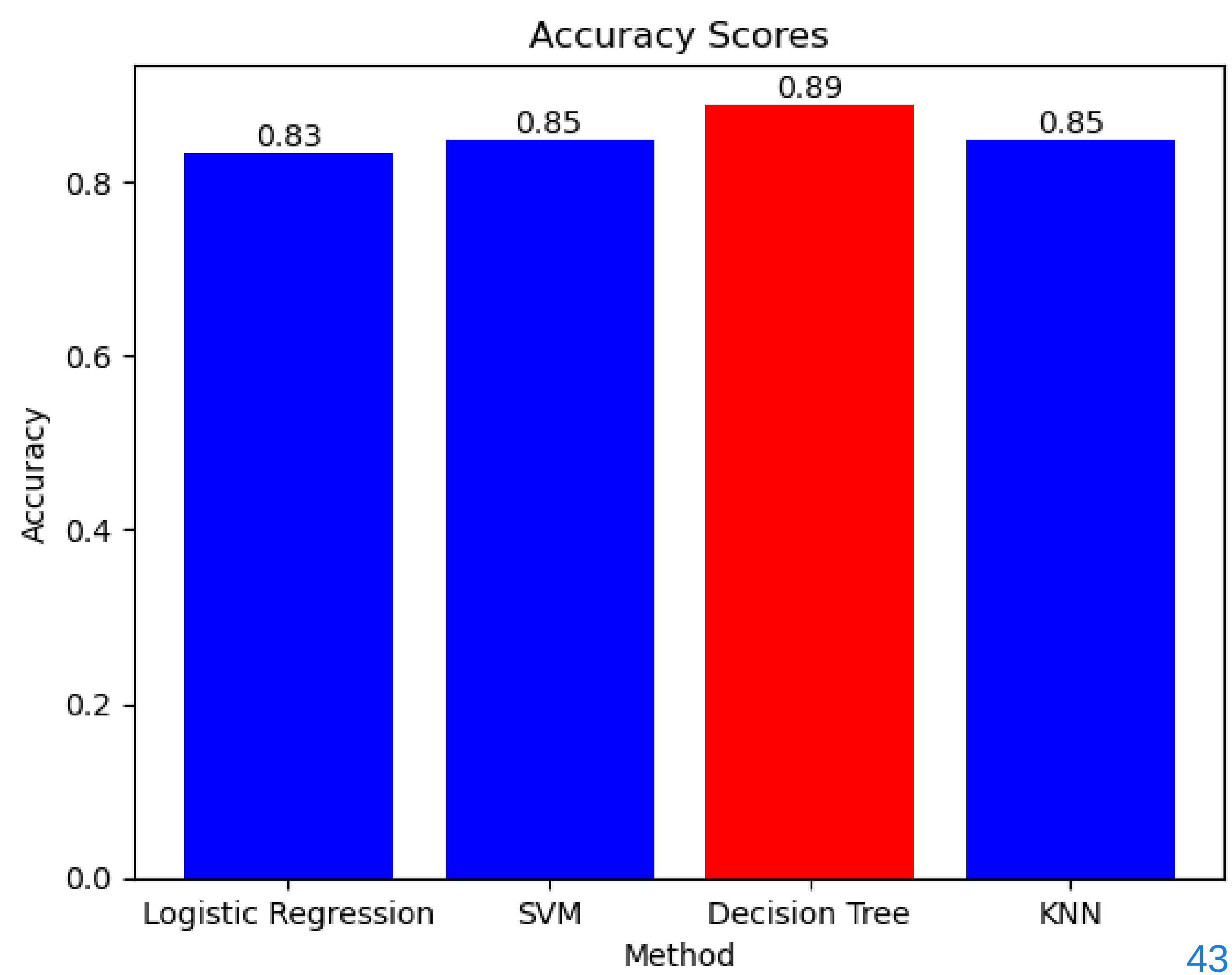


Section 5

Predictive Analysis (Classification)

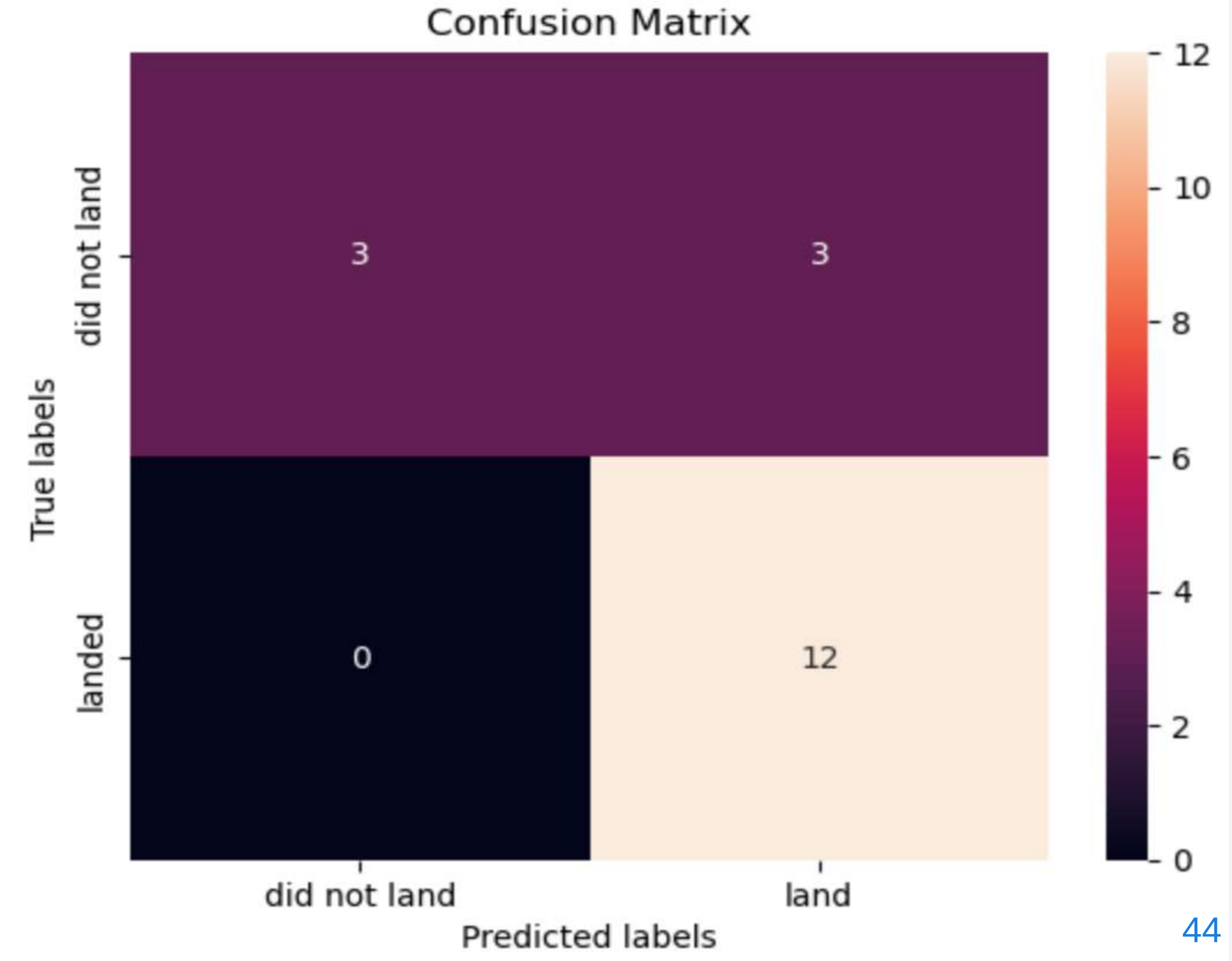
Classification Accuracy

- From par-chart, It can see that the Decision Tree model has the highest accuracy rate with 89%.



Confusion Matrix of the best model **(Decision Tree Model)**

- The confusion matrix for the Decision Tree model indicates that the model able to predict 12 success lands and 3 unsuccess lands which in line with true landing matrix. Where the model fail to predicate 3 unsuccess landings.



Conclusions

- The first 4 orbit types (ES-L1, GEO, HEO, SSO) had the best successful rate.
- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- In 2013 the success rate were gradually increased.
- 42% of the success launch was on the site (KSC LC- 39A) followed by (CCAFS LC- 40) with 29%.
- The highest total success launch rate which is in the site (KSC LC-39A) with 77% success launches.
- The FT booster version category has the highest success rate specially in the payload range of 2000-6000kg.
- Decision Tree model has the highest accuracy rate with 89%.

Thank you!